DOCUMENT RESUME

ED 465 790 TM 034 175

AUTHOR Weber, Deborah A.
TITLE Constructing Confidence Intervals for Reliability Coefficients Using Central and Noncentral Distributions.
PUB DATE 2002-04-03
NOTE 33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Reliability; *Research Reports; *Statistical Distributions
IDENTIFIERS *Confidence Intervals (Statistics)

ABSTRACT

Greater understanding and use of confidence intervals is central to changes in statistical practice (G. Cumming and S. Finch, 2001). Reliability coefficients and confidence intervals for reliability coefficients can be computed using a variety of methods. Estimating confidence intervals includes both central and noncentral distribution approaches. Some editorial guidelines now require prospective authors to report confidence intervals for score reliability coefficients (X. Fan and B. Thompson, 2001). Such requests follow the American Psychological Association Task Force on Statistical Inference recommendations for statistical methods in psychology journals and the American Psychological Association (2001) statement that confidence intervals are the best reporting strategy and strongly recommended. The paper illustrates methods of constructing confidence intervals. Three appendixes contain software text files for constructing confidence intervals. (Contains 2 tables and 31 references.) (Author/SLD)

Running head: CONFIDENCE INTERVALS FOR RELIABILITY COEFFICIENTS

Constructing Confidence Intervals for Reliability Coefficients Using Central and Noncentral Distributions

Deborah A. Weber

Texas A&M University 77843-4225

Paper presented at the annual meeting of the National Council On Measurement in Education, New Orleans, April 3, 2002.

Abstract

Greater understanding and use of confidence intervals is central to changes in statistical practice (Cumming & Finch, 2001). Reliability coefficients and confidence intervals for reliability coefficients can be computed using a variety of methods. Estimating confidence intervals includes both central and noncentral distribution approaches. Some editorial guidelines now require prospective authors to report confidence intervals for score reliability coefficients (Fan & Thompson, 2001). Such requests follow the American Psychological Association (APA) Task Force on Statistical Inference (TFSI) recommendations for statistical methods in Psychology journals and the American Psychological Association (2001) statement that confidence intervals are "the best reporting strategy"...and "strongly recommended" (p. 22, emphasis added).

Constructing Confidence Intervals for Reliability Coefficients Using Central and Noncentral Distributions

Editorial guidelines for prospective authors include reporting "confidence intervals for reliability coefficients whenever they report score reliabilities and note what interval estimation methods they have used" (Fan & Thompson, 2001, p. 517). Such requests follow the American Psychological Association (APA) Task Force on Statistical Inference (TFSI) recommendations for statistical methods in Psychology journals. Greater use of confidence intervals is part of the reform in behavioral and social sciences statistical practices (Cumming & Finch, 2001). According to Cumming and Finch (2001), four reasons why wider use of confidence intervals are promoted is because they: 1) are easily interpretable, 2) are connected to statistical significance tests, which are familiar to most people 3) promote meta-analytic thinking, and 4) give information regarding accuracy and precision. Estimating confidence intervals can be done in a variety of ways and includes both central and noncentral distribution approaches for various statistics including ANOVA (F, Fixed-effect, & Random-effect approaches), Cronbach's Coefficient Alpha, Person Correlation Coefficients, and the Split-Half Approach.

According to Fan and Thompson (2001), requesting that authors report confidence intervals for reliability coefficients furthers the goal of many editorial guidelines to "facilitate the movement of the field toward informed practices" (p. 218). According to the recommendations of Wilkinson and the APA Task Force (1999), "it is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval" (p. 599). Furthermore, Wilkinson and the APA Task Force (1999) addresses reliability by noting:

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of scores on a test for a particular population of examinees... Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores. (p. 596)

Additionally, Wilkinson and the TFSI (1999) stated "in all figures, include graphical representations of interval estimates whenever possible (p. 601). Furthermore, the American Psychological Association (2001) states in the 5th edition of the publication manual of the association, "because confidence intervals combine information on location and precision and can be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore, strongly recommended" (p. 22, emphasis added).

According to Fan and Thompson (2001), many authors use coefficients from previous studies because they are unaware that the reliability has changed. Many doctoral students and authors fail to understand measurement concepts that are fundamentally important to statistics and research design. As Thompson and Vacha-Haase (2000) emphasized, "ignorance about measurement realities creates a self-perpetuating resistance to overcoming misconception" (p. 180). According to Fan and Thompson (2001), it is important that readers realize all results are affected by sampling error variance and "reliability is no immutable property stamped into tests during the production process" (p. 219). Researchers should remember that reliability is not a property of the test but affected by the sampling error variance whenever the test is administered to a particular group (Thompson & Vacha-Haase, 2000; Vacha-Haase, Kogan, & Thompson, 2000). Despite effects by the APA Task Force and editorial guidelines designed to bring

awareness to problems with some commonly used current statistical methods, numerous researchers seem unfamiliar with many of the validity and reliability problems related to measurement concepts fundamentally important to research design.

According to a study by Whittington (1998), "problems related to validity and reliability of data-gathering procedures continue to exist in educational research literature, at least among the more selective research journals in the field of education" (p. 33). Whittington found questionable practices in at least one fifth of all the articles. Most notably was the lack of researchers reporting reliability information and failure to take into consideration population sample characteristics.

Unfortunately, the traditional mindset of statistical significance testing in psychology has limited the awareness of confidence intervals (CIs) by psychologists. As Cumming and Finch (2001) emphasized, in "psychological research, there is very little or no use of CIs and much room for improvement of statistical practices" (p. 536). The reasons for reporting confidence intervals is well-known according to Algina and Moulder (2001), who assert "reporting a point estimate alone fails to provide any indication of the uncertainty in the estimate" (p. 634).

The Need for Reporting CIs

Many researchers firmly endorse requests for reform (e.g. Cumming & Finch, 2001; Fan & Thompson, 2001; Smithson, 2001; Thompson, 1994; Thompson & Vacha-Haase, 2000; Vacha-Haase, et al., 1999) and are convinced that greater understanding and use of confidence intervals is central to changes in statistical practice within education, psychology, and related disciplines. Yet, despite recommendations by the Task force and criticism of statistical significance testing, a large number of psychologists' still fail to report CIs and/or effect sizes (Kieffer, Reese, & Thompson, 2001). According to Schmidt and Hunter (1997),

> Accepting the proposition that significance testing should be discontinued and replaced by point estimates and confidence intervals entrails the difficult effort of changing the beliefs and practices of a lifetime. Naturally such a prospect provokes resistance. Researchers would like to believe there is a legitimate rationale for refusing to make such a change. (p. 49)

Confidence intervals are not new to the field of statistics and reporting CIs promotes integration of results from previous research and informed judgments in relation to accuracy of estimates across studies (Fidler & Thompson, 2001). According to Wilkinson and the APA Task Force (1999) "comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies" (p. 599). Confidence intervals are easily comprehensible and provide information regarding the uncertainty of an estimation. The width of the confidence interval represents imprecision which typically includes the probability level, sampling error and error of measurement. For many, the failure to report confidence intervals is the result of uncertainty regarding the calculation of CIs or embarrassment about revealing the imprecision of many psychological studies (Cumming & Finch, 2001).

Definition of a Confidence Interval

Statements such as, support for the President was 90% with an error margin of 3%, gives the public information regarding a point estimate and uncertainty regarding the precision of the estimation that has been made. Such information is easy to understand and refers to confidence intervals for some population of interest. Despite contrary statements in many textbooks, a confidence interval of 95% DOES NOT mean there is a 95% chance that the CI contains the parameter in question. According to Fidler and Thompson (2001), a more accurate definition is to "frame a CI as one interval from an infinite or at least large sample of CIs for a given

parameter in which 1-α % of the intervals would capture the population parameter" (p. 578). As stated by Tietjen (1986),

> A (1-α) 100 percent *confidence interval* is an interval for an unknown parameter θ constructed from a large sample in such a way that if the same method were used to construct a "large" number of such intervals from independent samples, (1-α) 100 percent of the intervals would contain the parameter θ. (P. 35, cf. Fidler & Thompson, 2001)

According to Cumming and Finch (2001), "a CI is a set of parameter values that are reasonably consistent with the sample data we have observed." (p. 533). For a CI of 95%, theoretically, 95% of all the CIs would capture the population parameter and 5% would not (Fidler & Thompson, 2001). As Fidler and Thompson (2001) emphasized, it is essential to comprehend that "the parameter estimate and the endpoints for a *single* CI are all influenced by sampling error and thus changes from sample to sample" (p. 578). The use of a CI allows the reader and the researcher to make statistical inferences using components that have practical meaning to both. The inaccuracy indicated by the width of the CI results from a range of sources including error of measurement and sampling error. The confidence interval's width is dependent upon the probability level chosen, C. Wider CI's are typically the result of higher probability levels.

> It is never certain (unless we use C=100) that the interval includes the true values of the parameter of interest. The confidence interval or probability describes the chance of intervals of this kind including or "capturing" the population value it the long run. (Cumming & Finch, 2001, p. 533)

According to Fidler and Thompson (2001, p. 579), "CI's are typically computed by adding and subtracting from a given parameter estimate the standard error (SE) of that estimate times some

α or $\alpha/2$ centile of a relevant test distribution (e.g., t, F)". Additionally, CI are not to be interpreted with reference to zero (e.g. Thompson, 1998) but in comparison to prior studies as the point estimate (e.g. Schmidt, 1996). As Thompson (1998) postulated,

> if we interpret the confidence intervals in our study in the context of the intervals in all related previous studies, the true population parameters will eventually be estimated across studies, even if our prior expectations regarding the parameters are wildly wrong. (p. 800)

It is not in isolation, but through the interpretation of results in the context of previous studies that meta-analytical thinking is promoted.

According to Fan and Thompson (2001), many authors incorrectly use reliability coefficients from previous studies with the erroneous assumption that reliability is a property of the test, not the scores and therefore, unchanging. To further exacerbate the problem, students are no longer taught measurement concepts in many doctoral programs. As Thompson and Vacha-Haase (2000) stated,

> Ignorance about measurement realities creates a self-perpetuating resistance to overcoming misconceptions. We have entered a black-box era in which students with terminal degrees in education and psychology first enter their training based upon scores from a computer-adaptive GRE testing that upon their graduation they could not intelligently explain or evaluate. (p. 180)

According to Fan and Thompson (2001), reporting confidence intervals for reliability coefficients will enable readers to realize "all results within the general linear model, including measurement partitions of observed score variances, are influenced by sampling error

variances...[and] reliability is not an immutable property stamped into tests during the production process" (p. 519).

Estimating a Confidence Interval

When the parameter of μ is unknown, it can be estimated using an observed statistic (i.e. $\bar{x}$) to construct an interval using the observed statistic. According to Hinkle, Wiersma, and Jurs (1998, p. 219), a confidence interval is computed as,

CI = statistic ± (critical value) (standard error of the statistic)

For example, when the variance is known, the formula for the confidence interval for the mean becomes

$$CI = \bar{X} \pm (z_{cv})(\sigma_{\bar{x}})$$

Where $\bar{X}$ = sample mean

z_{cv} = critical value using the normal distribution

$\sigma_{\bar{x}}$ = standard error of the mean ($\sigma/\sqrt{n}$)

The critical value is determined by choosing a level of confidence. This is typically ($1-\alpha$) or $1-\alpha/2$ for a symmetric confidence interval. For example, a .01 level of significance (α) would have the corresponding level of confidence for construction of the confidence interval: 1 - .01 = .99. The value can then be determined by using a table containing areas under standard normal curve values of z (e.g. Hinkel et. al, 1998, p. 633). For a two tailed test of .01 level of significance we would divide .01 by 2 for a symmetric confidence interval. Therefore, .01/2 = .0050 and the closest z score of 2.57 would be used. The formula for the .99 percent confidence interval then becomes,

$$CI_{99} = \bar{X} \pm (2.57)(\sigma_{\bar{x}})$$

For this example, suppose the mean score of all students (N = 200, SD = 10) on a math test was 76.5. The 99 percent confidence interval is computed as follows:

$$CI_{99} = \bar{X} \pm (2.57)(\sigma_{\bar{x}})$$

$$= 76.5 \pm (2.57)(10/\sqrt{200})$$

$$= 76.5 \pm (2.57)(.71)$$

$$= 76.5 \pm 1.82$$

$$= (74.68, 77.32)$$

Remember, the width of the interval depends upon the probability level that is chosen. If everything is held equal, the higher the probability level, the wider the interval. For example, suppose we used the previous example to compute confidence intervals using the .05 level of significance. In this example, the confidence interval is 1 - .05 = .95. Once again the critical value is determined using a table containing areas under standard normal curve values of z. For a two tailed test we would once again divide .05 by 2 for a corresponding area beyond z of .0250. For this confidence interval:

$$CI_{95} = \bar{X} \pm (1.96)(\sigma_{\bar{x}})$$

$$= 76.5 \pm (1.96)(10/\sqrt{200})$$

$$= 76.5 \pm (1.96)(.71)$$

$$= 76.5 \pm 1.39$$

$$= (75.11, 77.89)$$

As you can see, the confidence interval is wider for the higher percentage of accuracy. In the previous examples, confidence intervals were computed for samples in which the population variance was known. However, we typically must use the variance of the sample to estimate the variance of the population (Hinkle et al., 1998). When this is the case, we do not use the normal

distribution when computing confidence intervals, but instead use the critical values from *t* distributions and the estimated standard error of the mean. According to Hinkel et al., 1998 (p. 220-221), the formula then becomes:

$$CI = \bar{X} \pm (t_{cv})(s_{\bar{x}})$$

where

$\bar{X}$ = sample mean

t_{cv} = critical value using the appropriate t distribution

$s_{\bar{x}}$ = estimated standard error of the mean ($s/\sqrt{n}$)

For example, if the sample mean ($\bar{X}$) for the 200 students is 76.5 and the estimated standard error ($s_{\bar{x}}$) is 10, the critical value is found for the confidence interval the same way the critical value for the test statistic was derived (Hinkle et al., 1998). To compute a 99-percent confidence interval we use the t distribution for n-1 = 200 = 199 degrees of freedom (since the variance is unknown). Using a table for the critical values of the t distribution (e.g. Hinkle et al., 1998) for a .01 level of significance, the corresponding critical value is found to be 2.57 for a two-tailed test. By applying the previous formula we get the following confidence intervals,

$$CI_{99} = \bar{X} \pm (t_{cv})(s_{\bar{x}})$$

$$= 76.5 \pm (2.57)(10/\sqrt{200})$$

$$= 76.5 \pm (2.57)(.71)$$

$$= 76.5 \pm 1.82$$

$$= (74.68, 77.32)$$

Similarly, if we want to compute the 95 = percent confidence interval, using 199 degrees of freedom we find the critical value is 1.96 for .05 level of significance. Therefore,

$CI_{95} = \bar{X} \pm (t_{cv})(s_{\bar{x}})$

$= 76.5 \pm (1.96)(10/\sqrt{200})$

$= 76.5 \pm (1.96)(.71)$

$= 76.5 \pm 1.39$

$= (75.11, 77.89)$

As Hinkle et al. (1998) emphasizes, "when we talk about the probability related to confidence intervals, we are talking about the probability that the confidence intervals constructed *from all possible samples of a given size* for a specific population will include μ" (p. 222).

Confidence Intervals and Null Hypothesis Significance Testing

According to Kristof (1963), there is an "intrinsic duality between hypotheses testing and constructing confidence intervals... (p. 236). As Hinkle et al. (1998) emphasized, both null hypothesis testing and computing confidence intervals involves the use of critical values, sample means, and standard error of the mean. For example, testing and rejecting the following null hypothesis at a .05 level of significance can also be done using confidence intervals.

where:

$H_o: \mu = 60.5$

$H_a: \mu \neq 60.5$

$\bar{X} = 76.5$ critical values $= \pm 1.96$ $\sigma_{\bar{x}} = .71$

we can compute the 95- percent confidence interval as:

$CI_{95} = \bar{X} \pm (1.96)(\sigma_{\bar{x}})$

$= 76.5 \pm (1.96)(10/\sqrt{200})$

$= 76.5 \pm (1.96)(.71)$

$= 76.5 \pm 1.39$

= (75.11, 77.89)

Using the confidence interval we can conclude that the mean of 60.5 is not contained with the confidence interval (75.11, 77.89). Since the hypothesized population value is NOT contained within the interval, we reject the null. By rejecting the null, we do not consider $H_o : \mu = 60.5$ a justifiable value and do not expect it in the interval. Confidence intervals also allow you to test several hypotheses at once. For example,

$H_o : \mu = 75.9$ $H_o : \mu = 76.5$ $H_o : \mu = 77.0$ $H_o : \mu = 77.8$

the above hypotheses would NOT be rejected because the values for each of the μ is contained within the confidence interval. Conversely,

$H_o : \mu = 60.5$ $H_o : \mu = 65.0$ $H_o : \mu = 70.5$ $H_o : \mu = 75.0$

the above hypotheses would be rejected because the values for each of the μ is not found within the confidence interval. In summary, "any value within the interval is a tenable value for the population parameter. All values outside the interval are not tenable" (Hinkle et al., 1998, p. 224).

As Fan and Thompson (2001) emphasized, stating confidence intervals for reliability coefficients will reinforce awareness that all statistical estimates, are influenced by sampling error variance (p. 517). It is important to note that the difference between null hypothesis significance testing (NHST) and CI's is that you have to have a hypothesis with NHST and you don't with confidence intervals. Unlike NHST, confidence intervals are not restricted to situations where a hypotheses can be reasonably established (Hinkle et al., 1998).

According to Thompson (1994), researchers often use null hypothesis statistical significance test of zero magnitude to test reliability coefficients. However, this method is not judicious because saying a reliability coefficient is significantly different from zero is not useful because

statistical significance is possible with a low reliability coefficient if the sample size is large enough. In statistical significance testing, sample size has a direct impact on significance. For example, "a one-tailed statistical significance test of an r of roughly .94, even at the $\alpha = .01$ level of statistical significance, will be statistically significant with an n as small as 5! (Thompson, 1994, p. 844). Unfortunately, many researchers advocate the use of NHST in favor of CI's because they illogically interpret CI's as subsuming zero. Ableson (1997), commented on using a nill null hypotheses on statistical tests of measurement and asserted, "when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80ish" (p. 121). Thompson (1994) goes on to state, "statistical tests of such coefficients in a measurement context makes little sense. Either statistical significance tests using the [nill] null hypothesis of zero magnitude should be by-passed, or meaningful null hypotheses should be employed" (p. 844). In addition to the obvious advantages of confidence intervals over null hypothesis statistical significance testing, confidence intervals also yields information regarding the accuracy and precision of measurement.

The benefit of using confidence intervals and reliability coefficients over null hypothesis statistical significance testing has been argued by many researchers (e.g. Cumming & Finch, 2001; Fan & Thompson, 2001; Meehl, 1967; Oakes, 1986; Rozeboom, 1960; Schmidt, 1996; Smithson, 2001; Steiger & Fouladi, 1997; Vacha-Haase, et al., 1999; Wilkinson & APA Task Force on Statistical Inference, 1999). According to Steiger and Fouladi (1997), confidence intervals afford all the information found in significance testing in addition to providing information regarding how big an effect is. Meaningful interpretation of p values require greater information than is provided in statistical significance testing and researchers should use caution

when comparing p levels. Steiger and Fouladi (1997) give the following example to illustrate this point,

> suppose someone reports a p level of .001. This could be representative of a trivial population effect combined with a huge sample size, or a powerful population effect combined with a moderate sample size, or a huge population effect with a small sample. (p. 226).

In addition to all the information provided by the p level, the width of the confidence interval provides information regarding the precision of measurement that is not available with statistical significance testing. Reporting confidence intervals affords a superior alternative to the traditional null hypothesis testing.

Reporting CIs for Reliability Coefficients

There are various ways to estimate confidence intervals for reliability coefficients. According to Educational and Psychology Measurement,

> EMP authors should report confidence intervals for reliability estimates whenever they report score reliabilities and note what interval estimation methods they have used. This will reinforce reading understanding that all statistical estimates, including those for score reliability, are affected by sampling error variance. (Fan & Thompson, 2001, p. 517)

Constructing CIs with the appropriate distributions include both central distributions (central t) and noncentral distributions (noncentral t). According to Cumming and Finch (2001), most inferential techniques use central distributions (e.g., t distribution), however, non-zero centered distributions (e.g. noncentral distributions) are also an important part of inferential statistics. As Fan and Thompson (2001) stated,

> In addition to being necessary for many powerful computations, noncentral distributions are also necessary for accurate calculation of results invoking ratios of estimates to other estimates. Thus, noncentral intervals are needed to compute accurate confidence intervals for standardized effect sizes. (p. 522)

While only one parameter is described in central t distributions (e.g. degrees of freedom), an additional parameter (Δ) is used in noncentral distributions. According to Cumming and Finch (2001),

> Central t distributions, which are always symmetric, arise when a normally distributed variable with a mean of zero is divided by an independent variable closely related to the χ^2 distribution. Noncentral t distributions arise when a normally distributed variable with mean *not* equal to zero is divided by an independent variable closely related to the χ^2 distribution. They are not symmetric and the degree to which they are skewed depends on Δ, the distance by which the mean of the normal distribution is displaced from zero. (p. 547)

Despite their usefulness in statistical inference, many researchers remain unaware of noncentral test statistic distributions, which, until recently, have been impractical due to the lack of computer software to make such estimates (Fidler & Thompson, 2001). However, as Fan and Thompson stated (2001), noncentral distributions may be even more relevant when obtaining CIs for reliability estimates due to the larger effect sizes expected in reliability studies.

Constructing CIs for reliability coefficients can encompass a variety of interval estimation methods and CIs can be computed using both central and noncentral distribution methods. ANOVA results can also be utilized to compute reliability coefficients and confidence intervals for score reliabilities using Fixed-effects ANOVA, Random Effects Approach and ANOVA F

values. Additionally, approaches such as Cronbach's Coefficient Alpha, Person Correlation Coefficient, the Split-Half Approach can all be utilized in reporting confidence intervals for reliability coefficients (Fan & Thompson, 2001).

ANOVA

ANOVA effects may be random or fixed. According to Fidler and Thompson (2001), an ANOVA effect is random if we "randomly select some levels of a way while maintaining the capacity to generalize to the population of levels" (p. 581). Additionally, a random effect design is used when at least one random-effect way and one fixed-effect way are part of a muliway case. A fixed effect is when all the levels of a way are used or certain levels of a way are used but we do not generalize outside the chosen levels. Mixed-effects, random effects, and fixed effects models are all calculated the same way (i.e. Sum of Squares, Mean Squares). However, different denominators are utilized when computing variance and F values, and consequently generalizations across the different models differ depending on what denominator was used in the F tests calculation. Fidler and Thompson (2001) make this distinction clear,

> For example, for a two-way fixed-effects factorial design, all effect Mean Squares are divided by Mean $\text{Square}_{\text{ERROR}}$. But when both ways involve random effects, the main effect F tests are instead computed using the Mean $\text{Square}_{\text{INTERACTION}}$ as the denominator of the calculation, whereas the interaction F test is still computed using the Mean $\text{Square}_{\text{ERROR}}$ as the denominator. (p. 581)

Numerous effect sizes can be calculated in every study.

[Insert Table 1 here]

Using the heuristic data in Table 1, computing a reliability coefficient using ANOVA results can be done using Hoyt's methods (Fan & Thompson, 2001) where,

$$(MS_{people} - MS_{error}) / MS_{people}$$

$$(.4737 - .894) / 4.737$$

$$3.843 / 4.737 = 0.8113$$

[Insert Table 2 here]

Estimations using variance components can also be calculated using the following formula (Fan & Thompson, 2001, p. 521),

$$[V_{people} / n_p] / [(V_{people} / n_p) + (V_{p*v,e}) / n_p n_v)]$$

$$[1.281/15] / [(1.281 / 15) + (.894 / 15 (3))]$$

$$.0845 / [.0845 + (0.894 / 45)]$$

$$.0854 / .0845 + 0.019866666$$

$$.0854 / 0.105266666 = .8113$$

Cronbach's Coefficient Alpha

Cronbach's coefficient alpha can also be used for reliability computations. Using the data in Table 2, Cronbach's coefficient alpha can be computed as follows,

$$\text{Cronbach's } \alpha = [k / (k - 1)] [1- (\Sigma V_k / V_{total})$$

$$[3 / 2] [1 - ((2.2667 + 1.8571 + 2.4000) / 14.210)]$$

$$[1.5] [1 - (6.5238 / 14.210)]$$

$$[1.5] [1 - .4591]$$

$$[1.5] [.5409] = .8113$$

According to Fan and Thompson (2001), ANOVA F values can be used to obtain confidence intervals for reliability coefficients involving relative decisions (the interested reader should see Burdick & Graybill, 1992 for information on computing CIs for absolute decisions). It is in this context that reliability coefficients for central approaches, such as Cronbach's alpha

(Charter & Feldt, 1996; Feldt, 1965, 1980, 1990; Feldt, Woodruff, & Salih, 1987; Kristof, 1963) are applicable. According to Fan and Thompson (2001),

> for a given sample of n examinees taking a test with k items, the upper and lower confidence interval limits for the sample Cronbach coefficient alpha α a the given statistical significance level γ can be constructed as
>
> $CI_{upper} = 1-[(1-\alpha) \text{ X } F_{(\gamma/2),df_1,df_2}]$, and
>
> $CI_{lower} = 1-[(1-\alpha) \text{ X } F_{(1-\gamma/2),df_1,df_2}]$
>
> where F represents the values of the F distribution for percentiles $\gamma/2$ and $1-\gamma/2$, respectively, with $df_1 = (n-1)$ and $df_2 = (n-1)(k-1)$. (p. 522)

Using our heuristic data, confidence intervals can be computed by,

$df_1 = 15 - 1 = 14$

$df_2 = (15 - 1)(3-1) = 28$

$F_{(.025, 14, 28)} = .3635$ (lower percentile F for the CI_{upper} using Excel command for F value), "$=FINV_{(.025, 14, 28)}$".

$F_{(.975, 14, 28)} = 2.374$ (upper percentile F for the CI_{lower} using Excel command for F value), "$=FINV_{(.975, 14, 18)}$".

$$CI_{upper} = 1-[(1-\alpha) \text{ X } F_{(\gamma/2),df_1,df_2}]$$

$$1-[(1 - .8113) \text{ X } .3635)]$$

$$1-[0.1887 \text{ X}.3635)]$$

$$1 - 0.0686 = .9314$$

$$CI_{lower} = 1-[(1-\alpha) \text{ X } F_{(1-\gamma/2),df_1,df_2}]$$

$$1 - [(1-.8113) \text{ X } 2.374]$$

$$1 - [0.1887 \text{ x } 2.374]$$

$$1 - 0.448 = .5520$$

Additionally, Fan and Thompson (2001, p. 524) supply the syntax for obtaining reliability coefficients using SPSS as follows,

```
reliability variables=v1 to v3/
    scale (TOTAL)=v1 to v3/
    statistics=corr cov/summary=means var total/
    icc=model(random) type(consistency) cin=95 testval=.70/
    model=alpha
```

The heuristic example in appendix A has a reliability of .8113. The reliability coefficient computed using Hoyt's methods, variance components, and Cronbach's coefficient alpha (Fan & Thompson, 2001) all resulted in a reliability coefficient of .8113, the same reliability coefficient obtained through the SPSS syntax in appendix C.

Fixed-Effects ANOVA

Computing a confidence interval for a fixed-effect ANOVA can also be done using SPSS. According to Smithson (2001), "when working with statistics for which we deem a one-sided interval appropriate, 100(1-α)% CIs are computed by declaring 'CONF' to be 1-[2(α)] (e.g., use 'CONF'=.90 to obtain a 95% one-sided interval)" (p. 613). 'CONF' is declared to be 1-α (.95), if we want the interval to be two-sided. Since reliability is estimated as an unsquared value we use the square root of the confidence interval boundaries to obtain the interval (e.g. see Smithson, 2001 for complete SPSS syntax).

Random-Effects ANOVA

According to Fan and Thompson (2001), confidence intervals for reliability coefficients can be estimated using the "R2" computer program or using a regression logic (Fan & Thompson,

2001). For the R2 value we would input the reliability coefficients and n = df1 + df2 + 1. The number of predictor variables plus the single criterion variable equals the number of degrees of freedom numerator. Once the confidence interval is obtained the square roots are taken to obtain the boundaries.

Pearson Correlation Coefficients

According to Fan and Thompson (2001), test-retest and interrater reliability estimates are all Pearson correlation coefficients and confidence intervals can be computed using the following four steps:

1. transform r to Z_r. (Fisher Z transformation);
2. compute σ_z: $\sigma_z = 1/(n-3)^{.5}$ (n = number of examiners, or raters)
3. obtain CI for Z_ρ: $Z_r \pm 1.96\sigma_z$ (for 95% CI) and
4. transform lower/upper limits back to Pearson r. (p. 525)

Additionally, these calculations can be constructed on SPSS, SAS, or a spreadsheet program.

Split-Half Approach

Using the Spearman-Brown prophecy formula, a correlation coefficient can be obtained for the entire test. Using the previously described four steps, a confidence interval can be constructed for the correlation coefficient between the two halves of the test. Once the confidence limits are obtained, the Spearman-Brown formula is applied to obtain the reliability estimate for the entire test.

Summary

Reliability coefficients and confidence intervals for reliability coefficients can be computed using a variety of methods. Estimating confidence intervals includes both central and noncentral

distribution approaches. While authors retain the freedom to choose-from numerous estimation method, it is important to note which method has been used when reporting score reliabilities.

Greater understanding and use of confidence intervals is central to changes in statistical practice. As Cumming and Finch (2001) emphasized, in "psychological research, there is very little or no use of CIs and much room for improvement of statistical practices" (p. 536). Thompson (2002) provides a detailed explanation f the appropriate use and interpretation of confidence intervals in contemporary research.

Despite criticism of statistical significance testing and recommendations by the Task Force and the American Psychological Association (2001) statement that confidence intervals are "the best reporting strategy"...and "strongly recommended" (p. 22, emphasis added), a large number of psychologists' still fail to report CIs and/or effect sizes (Kieffer et al., 2001). According to Cumming and Finch (2001), statistical practices in behavioral and social sciences will require greater use of meta-analysis, effect size measures, and confidence interval before any highly desirable reform occurs within the field. Additionally, wider use of confidence intervals is promoted because CI's are easily interpretable, are connected to statistical significance tests, which are familiar most, promote meta-analytic thinking, and give information regarding accuracy and precision.

Table 1

Heuristic Data

	Item			
Person	1	2	3	Total
1	6	6	7	19
2	9	10	9	28
3	8	10	7	25
4	8	10	9	27
5	9	11	10	30
6	12	10	11	33
7	10	9	9	27
8	9	10	10	29
9	9	8	10	27
10	7	7	7	20
11	7	8	9	24
12	10	9	8	27
13	8	10	9	27
14	8	9	10	27
15	7	8	6	21
Mean	8.47	9.00	8.60	26.0667
Variance	2.267	1.857	2.400	14.210

Table 2

Source		Sum of Squares	df	Mean Squares
INTERCEPT	Hypothesis	3397.356	1	3397.356
	Error	54.321	10.869	4.998[a]
PERSON	Hypothesis	66.311	14	4.737
	Error	25.022	28	.894
ITEM	Hypothesis	2.311	2	1.156
	Error	25.022	28	.894[b]

Appendix A

ASC TEXT Data File for ANOVA Estimates of Score Reliabilities

```
1 1 6
1 2 6
1 3 7
2 1 9
2 2 10
2 3 9
3 1 8
3 2 10
3 3 7
4 1 8
4 2 10
4 3 9
5 1 9
5 2 11
5 3 10
6 1 12
6 2 10
6 3 11
7 1 10
7 2 9
7 3 8
8 1 9
8 2 10
8 3 10
9 1 9
9 2 8
9 3 10
10 1 7
10 2 7
10 3 6
11 1 7
11 2 8
11 3 9
12 1 10
12 2 9
12 3 8
13 1 8
13 2 10
13 3 9
14 1 8
14 2 9
14 3 10
```

15 1 7
15 2 8
15 3 6

Appendix B

ASC Text Data File for Estimates of Variance Components of Score Reliabilities

```
6  6  7
9  10 9
8  10 7
8  10 9
9  11 10
12 10 11
10 9  8
9  10 10
9  8  10
7  7  6
7  8  9
10 9  8
8  10 9
8  9  10
7  8  6
```

Appendix C

SPSS Syntax File for ANOVA and Variance Components Estimates of Score Reliability

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=listing.
TITLE 'Deborah Weber Reliability Coefficients (2002)' .
DATA LIST
File='a:\ASCTEXTDATA.txt' FIXED RECORDS=1 TABLE/1
Person 1-2 Item 4 Score 6-7.
list variables=all/cases=9999/format=numbered.

SubTitle 'RANDOM EFFECTS MODELS' .
execute.
UNIANOVA
        score BY person item
        /RANDOM = person item
        /METHOD = SSTYPE(3)
        /INTERCEPT = INCLUDE
        /PRINT = DESCRIPTIVE
        /CRITERIA = ALPA(.05)
        /DESIGN = person item.
VARCOM
        score BY person item
        /RANDOM = person item
        /METHOD = SSTYPE(3)
        /DESIGN = person item
         /INTERCEPT = INCLUDE .

DATA LIST
File='a:\ASCTEXTDATA2.txt' FIXED RECORDS=1 TABLE/1
Var1 1-2 Var2 4-5 Var3 7-8.

compute total=sum(var1 to var3) .
list variables=all/cases=9999/format=numbered.
descriptives variables=all/statistics=all .
SubTitle 'CONFIDENCE INTERVALS'.
execute.

reliability variables=Var1 to Var3/
        scale(TOTAL)=Var1 to Var3/
        statistics=corr cov/summary=means var total/
        icc=model(random) type(consistency) cin=95 testval=.70/
        model=alpha.
```

References

Abelson, R.P. (1997). A retrospective on significance test ban of 1999 (If there were no significance tests, they would be invented). In L. Harlow, S. Muliak, & J. Steiger (Eds.), *What if there were no significance tests*? (pp. 117-141). Mahwah, NJ: Lawrence Erlbaum.

Algina, J. & Moulder, B.C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement, 61*(4), 633-649.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Burdick, R.K., & Graybill, F.A. (1992). *Confidence intervals for variance componen*ts. New York: Marcel Kedder.

Cumming, G., & Finch, S. (2001). A primer on understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532-574.

Charter, R.A., & Feldt, L.S. (1996). Testing the equality of two alpha coefficients. *Perceptual and motor skills, 82*, 763-768.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological measurement, 61*(4), 517-531.

Feldt, L.S. (1965). The approximate sampling distribution of kuder-richardson reliability coefficient twenty. *Psychometrika, 30*(3), 357-370.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99-105.

Feldt, L.S. (1990). The sampling theory for the interclass reliability coefficient. *Applied*

Measurement in Education, 3, 361-367.

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. *Applied Psychological measurement, 11*(1), 93-103.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed -and random effects sizes. *Educational and Psychological Measurement, 61*, 575-604.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavior sciences* (4th ed.). Boston: Houghton Mifflin.

Kieffer, K.M., Reese, R.J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69*, 280-309.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*, 221-238.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Oakes, M.L. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: John Wiley.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significant test. *Psychological Bulletin, 57*, 416-428.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implication for the training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Muliak, & J. Steiger (Eds.), *What if there were no significance tests*? (pp. 37-64). Mahwah, NJ: Lawrence

Erlbaum.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61*, 605-632

Steiger, J.H., & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What If There Were No Significance Tests*? (pp. 221- 257). Mahwah, NJ: Lawrence Erlbaum

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*(4), 837-847.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195.

Tietjen, G.L. (1986). *A topical dictionary of statistics*. New York: Chapman and Hall.

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*(4), 509-522).

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education, 67*(4), 335-341.

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in

psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement,58*(1), 21-37.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
NCME

REPRODUCTION RELEASE

TM034175

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Constructing Confidence Intervals for Reliability Coefficients Using Central And Noncentral Distributions	
Author(s): Deborah A. Weber	
Corporate Source: Texas A+M University	Publication Date: 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Deborah Weber	Printed Name/Position/Title: Deborah Weber, M.Ed
Organization/Address: 2230 Dartmouth # 1215	Telephone: 979-680-9753 FAX:
	E-Mail Address: deborahweber@tamu.edu Date: 5/17/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com